Inštrukcie k spusteniu procesu konsolidácie RA

Tento dokument obsahuje inštrukcie k spusteniu transformačného a konsolidačného modulu pre databázu Register Adries.

# Transformačný modul

Transformačný modul slúži na transformáciu jednotlivých inicializačných dávok z formátu XML do formátu JSON, v ktorom sa dáta posielajú do databázy Registra Adries. Transformáciu je potrebné vykonať pre každú novú inicializačnú dávku, ktorá je poskytnutá Ministerstvom Vnútra, a to ešte **pred** spustením konsolidačného modulu. Zdrojové inicializačné dávky vo formáte XML **musia** byť uložené v adresári s cestou *.../init\_batch/xml/* relatívne k adresáru s ostatnými zdrojovými súbormi, ktorého cesta bude poskytnutá ako argument pri spustení programu (viď nižšie).

# Konsolidačný modul

Konsolidačný modul slúži na obnovenie dátových zdrojov Registra Adries. Prebieha v nasledujúcich krokoch pre každý z ôsmich datasetov (registrov):

**Krok 1:** Stiahnutie a zazipovanie všetkých pôvodných dátových zdrojov.

**Krok 2:** Odstránenie tabuliek z datastore dátových zdrojov (zmenové a konsolidované dáta), kompletné vymazanie ostatných dátových zdrojov (dokumentácia a inicializačná dávka).

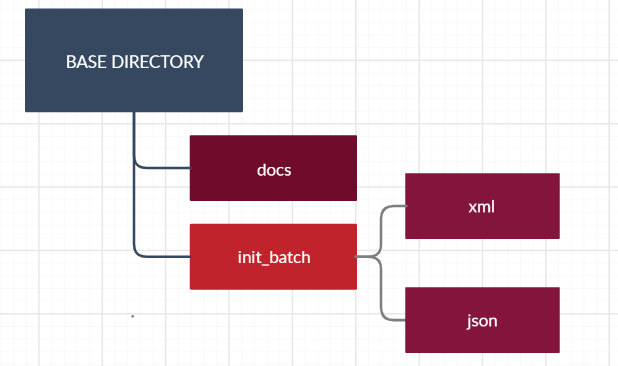
**Krok 3:** Upload nových dokumentácii a nových inicializačných dávok (v prípade, že je inicializačná dávka príliš veľká je počas behu programu zkomprimovaná a uploadnutá je jej .zip verzia).

**Krok 4:** Vytvorenie nových datastore tabuliek v existujúcich dátových zdrojoch (zmenové, konsolidované a init+zmenové dáta), následne upload novej inicializačnej dávky do konsolidovanej a init+zmenovej tabuľky.

**Krok 5:** Stiahnutie všetkých zmenových dávok, ktoré boli vypublikované od času kedy bola zhotovená nová inicializačná dávka (a teda ešte neboli aplikované) a ich upload do jednotlivých datastore tabuliek.

## Inštrukcie

Pre bezchybný beh programu je nutné dodržať nasledujúcu adresárovú štruktúru:



Kde *BASE\_DIRECTORY* je adresár, ktorého absolútna cesta je zadaná ako prvý vstupný argument programu (príklad cesty: */home/user/Desktop/register-adries/resources****/*** , *BASE\_DIRECTORY* je v tomto prípade adresár *resources*),

*init\_batch/xml/* obsahuje inicializačnú dávku vo formáte XML,

*init\_batch/json/* obsahuje adresáre s jednotlivými inicializačnými dávkami vo formáte JSON, ktoré boli dopredu vytvorené spustením transformačného modulu,

*docs/* je adresár obsahujúci dokumentácie jednotlivých datasetov vo formáte HTML (upravené dokumentácie pre všetkých 8 registrov budú poskytnuté spolu s ostatnými súbormi, keďže konsolidovaná tabuľka bude podľa žiadosti obsahovať navyše aj stĺpce, ktoré boli doteraz iba v zmenových dátach). Pri každom ďalšom behu stačí v tomto adresári ponechať jeho pôvodný obsah a jednotlivé dokumentácie sa tým nezmenia.

Názvy zdrojových súborov uložené v *init\_batch/xml/* a *doc/* musia mať nasledujúcu formu:

<názov\_súboru>**-0X.xml** pre XML inicializačné dávky

<názov\_súboru>**-0X.html** pre dokumentácie datasetov

kde **X** treba nahradiť poradovým číslom daného datasetu (**1** – Kraje, **2** – Okresy, **3** – Obce, **4** – Časti obcí, **5** – Ulice, **6** – Budovy, **7** – Vchody, **8** - Byty). Za <názov súboru> je možné doplniť čokoľvek.

*BASE\_DIRECTORY* musí taktiež obsahovať súbor **apikey.txt**, v ktorom je na jedinom riadku uložený platný autorizačný kľúč pre CKAN API pre databázu Registra Adries na data.gov.sk.

Celý beh transformačného aj konsolidačného modulu trvá niekoľko hodín (hlavne kvôli registrom Budovy a Vchody, ktoré obsahujú niekoľko miliónov záznamov). Pre úspešný beh celého procesu je z toho istého dôvodu potrebné mať voľné miesto na disku aspoň približne **6 GB** a zaistiť, že sa počas celého behu počítač neprepne do spánkového režimu alebo nevypne.

## Príklad spustenia modulov z príkazového riadku

1. Transformačný modul (**transformation.jar**)

*java -jar transformation.jar /home/user/Desktop/register-adries/resources/*

1. Konsolidačný modul (**register-adries-fix.jar**)

*java -jar register-adries-fix.jar /home/user/Desktop/register-adries/resources/ 0*

- 1. argument predstavuje absolútnu cestu k *BASE\_DIRECTORY* na danom PC

- 2. argument (voliteľný) je *changes\_id* poslednej zmenovej dávky, ktorá bola na aplikovaná pred vytvorením inicializačnej dávky. Všetky novšie zmenové dávky budú aplikované v rámci behu programu, a teda nedôjde k strate žiadnych záznamov. Ak nie je potrebné aplikovať žiadne zmenové dávky vynechajte tento argument. Keďže pri prvom behu sa používa inicializačná dávka z 06/2018 a bude treba aplikovať všetky zmenové dávky, tento argument by mal byť 0.

V prípade, že konsolidačný proces zlyhá medzi vymazaním obsahu pôvodných záznamov a nahraním nových v dôsledku nepredvídateľnej chyby, budú všetky pôvodné dáta z CKANu zálohované v priečinku *BASE\_DIRECTORY/backup/<názov registra pri ktorom nastala chyba>.*

Po **prvom úspešnom dobehnutí** **konsolidačného modulu je nutné** **neodkladne spustiť** aj priložený program **ra-ws-process-update.jar**, ktorý vykoná nevyhnutné zmeny a prispôsobenia na automatizovanom procese sťahovania a aplikovania zmenových dávok, ktorý sa spúšťa v pravidelných denných intervaloch. Program je potrebné spustiť s jediným argumentom nasledujúcim príkazom:

java -jar ra-ws-process-update.jar /…/ra\_ws\_muleprocess/

- kde jediný argument predstavuje absolútnu cestu ku koreňovému adresáru so zdrojovými súbormi k procesu *ra\_ws\_muleprocess* (ktorá bude pravdepodobne vyzerať nasledovne: */usr/share/mule/apps/ra\_ws\_muleprocess/)*.